Sven Leuckert

Typological Interference in Information Structure: The Case of Topicalization in Asia

Erstveröffentlichung in / First published in:

Zeitschrift für Anglistik und Amerikanistik. 2017, 65(3), S. 283– 302 [Zugriff am: 30.01.2020]. De Gruyter. ISSN 2196-4726

DOI: https://doi.org/10.1515/zaa-2017-0029

Sven Leuckert*

Typological Interference in Information Structure: The Case of Topicalization in Asia

DOI 10.1515/zaa-2017-0029

Abstract: Topicalization refers to the sentence-initial placement of constituents other than the subject and is often listed as a non-canonical construction [cf. Ward, Gregory, Betty J. Birner and Rodney Huddleston (2002). "Information Packaging." Rodney Huddleston and Geoffrey K. Pullum, eds. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, 1363–1447.]. In this paper, tokens of topicalization in the direct conversations in the *International Corpus of English* for Hong Kong and India and, for comparison, Great Britain are analysed. In order to find out if topicalization is a contact-induced feature, typological profiles with regard to topic-prominence [Li, Charles N. and Sandra A. Thompson (1976). "Subject and Topic: A New Typology of Language." Charles N. Li, ed. *Subject and Topic*. New York: Academic Press, 457–489.] are created for three Indo-Aryan, three Dravidian and two Sinitic languages. I suggest that the low frequencies of topicalization in Hong Kong English and the high frequencies of topicalization in Indian English are primarily due to differences in intensity of contact [Thomason, Sarah G. (2001). *Language Contact*. Washington, D.C.: Georgetown University Press.] and variety development [Schneider, Edgar W. (2007). *Postcolonial English. Varieties Around the World*. Cambridge: Cambridge University Press.]. Typological interference at the level of information structure is assumed to only come to the fore in further developed varieties and after prolonged contact.

1 Introduction

Topicalization, i.e. the sentence-initial placement of constituents other than the subject, is listed in the *Cambridge Grammar of the English Language* as a non-canonical construction (Ward et al. 2002, 1365). Compare the following two examples illustrating the 'canonical' version of a sentence (1) and its 'non-canonical' counterpart with topicalization (2) (Ward et al. 2002, 1365; emphasis added):

(1) We rejected six of the applications.
(2) *Six of the applications* we rejected.

***Corresponding author: Sven Leuckert,** Technische Universität Dresden, Institute of English and American Studies, Wiener Str. 48, 01069 Dresden, Germany, e-mail: sven.leuckert@tu-dresden.de

This study focuses on topicalization in non-native varieties of English, particularly in Asian Englishes. Topicalization has been noted as a feature of South African Indian English (henceforth SAIE, Mesthrie 1992), South African Black English (Mesthrie 1997), Indian English (Lange 2012) and several other L1 and L2 varieties of English (Winkle 2015). Examples (3) and (4) show the phenomenon as it may occur in non-native varieties, with the topicalized constituent printed in italics:

(3) *Linguistic* I'm going to quit
<ICE-HK:S1A-028#108:1:C>
(4) *Admissions* you are getting
<ICE-IND:S1A-040#239:1:A>

In the present paper, the frequency of topicalization in Indian English (IndE), Hong Kong English (HKE) and, for comparison, British English (BrE) is analysed based on the direct conversations in the *International Corpus of English* (ICE).[1] One of the main questions that requires in-depth research concerns the reasons behind the feature's high frequency in IndE and its low frequency in HKE. Since Li and Thompson's (1976) seminal paper on the distinction between subject and topic, certain languages have been claimed to be more 'topic-prominent' (e.g. Mandarin Chinese) and others more 'subject-prominent' (e.g. most Indo-European languages). Topic-prominence refers to the state of structuring sentences according to the topic-comment principle, i.e. placing what the sentence is about (the topic) at the beginning of the clause and the information about it (the comment) behind the topic. Rather than making claims about the presence or absence of subjects and topics in a language, however, Li and Thompson postulate that the overall configuration of a language may be shaped according to one of the two principles or a mixture of both. Since several Asian varieties of English have been in contact with languages that fall into the topic-prominent group, a contact-based explanation immediately appears attractive. The underlying assumption in this line of reasoning is that a tendency towards structuring sentences after a topic-comment model in lieu of the common English subject-predicate pattern is part of Asian speakers' linguistic repertoire, resulting in increased frequencies of topicalization – a canonical or unmarked phenomenon in highly topic-prominent languages – in their English sentences. However, the earliest of the sources listed above, namely Mesthrie's work on SAIE from the beginning of the 1990s, rejects language contact as an explanatory parameter for topicalization:

1 In the following, I will occasionally refer to the ICE components as ICE-GB (=Great Britain), ICE-HK (=Hong Kong) and ICE-IND (=India).

> The predilection for topicalisation […] is not substrate-induced. The Indic and Dravidian languages do not appear to use a particularly striking proportion of topicalised sentences (no more than standard English, say). Once again we see universals of discourse structure playing a greater role than transfer. (Mesthrie 1992, 157)

In more recent sources, scholars meet a contact hypothesis much more enthusiastically. Lange (2012, 151), for instance, calls for an analysis of contact languages in order to explain the forms and high frequencies of topicalization in spoken IndE: "Looking at all the evidence available, I would strongly suggest to take a closer look at substrate influence as the decisive factor for the form and frequency of topicalization constructions in spoken IndE."

Clearly, a close analysis of the typology of the major contact languages of Asian varieties with regard to topicalization appears to be a desideratum. In this paper, I provide an analysis of topic-prominence in some major Indo-Aryan, Dravidian and Sinitic languages in order to evaluate the potential of analysing topicalization as a contact-induced feature. Paying due respect to the fact that contact cannot be taken as the sole explanation, however, I also consider other factors and detail some of the theoretical aspects of language contact in relation to information structure. Furthermore, I will comment on the distinction between canonical and non-canonical with regard to my findings and identify in which way topicalization can be considered to be one or the other in the two Asian varieties. Finally, I will discuss the idea that the influence of substrate languages in the case of topicalization appears to be linked to the degree of nativisation, with the typological configuration of a contact language playing a role only at later stages of variety development.

In Section 2, I discuss definitions of the terms 'topicalization' and 'topic-prominence' and how these terms are related. Section 3 gives the results of my empirical analysis of topicalization in ICE-IND, ICE-HK and ICE-GB. Section 4 discusses the results, focusing on the role of language contact and variety status. The final section provides a conclusion and an outlook.

2 Topic-Prominence and Topicalization

The terminology applied to name and describe the sentence-initial placement of constituents other than the subject in English varies considerably, with terms such as 'preposing' (cf. Ward et al. 2002; Birner and Ward 2009), 'fronting' (Winkle 2015) and 'topicalization' (Mesthrie 1992; Lange 2012) used more or less interchangeably in the literature. An important theoretical framework has been established by Ward and Birner (1998) and several of their subsequent publications.

Based on a study by Ward (1988), Birner and Ward define preposings as constructions "in which a lexically governed phrasal constituent (NP, AP, PP, VP)

appears to the left of its canonical position, typically sentence-initially" (Birner and Ward 1998, 3; cf. also Ward and Birner 2001, 124–126; Ward et al. 2002, 1374; Ward and Birner 2004, 158; Birner and Ward 2009, 1172–1173; Ward and Birner 2011, 1938). They consider left-dislocation to be outside of the realm of preposing (cf. Birner and Ward 1998, 5), because unlike preposing constructions, left-dislocation allows for discourse-new and hearer-new initial constituents (cf. Ward and Birner 2004, 162). Furthermore, the syntax of the matrix clause in left-dislocation remains intact due to a resumptive pronoun standing in for the fronted constituent. Additionally, Ward and Birner differentiate between topicalization and focus preposing. Examples (5–7) illustrate cases of topicalization, focus preposing and left-dislocation (adapted from Ward and Birner 2004, 160–162; emphasis added):

(5) G: Do you watch football?
E: Yeah. *Baseball* I like a lot BETTER.
(6) Colonel Kadafy, you said you were planning on sending planes – *M16s* I believe they were – to Sudan.
(7) *One of the guys I work with, he* said he bought over $100 in Powerball tickets.[2]

Topicalization 'proper' (5) and focus preposings (6), according to Ward and Birner, differ in terms of their information status and their intonation. They note that

> [t]he focus in a topicalization […] is not contained in the preposed constituent but occurs elsewhere in the utterance. Intonationally, preposings of this type contain multiple accented syllables: (at least) one occurs within the constituent that contains the focus and (at least) one occurs within the preposed constituent […]. (Ward and Birner 2004, 161)

For the present analysis, I did not distinguish between topicalization in the narrow sense and focus preposing. This is due to the fact that no audio files are readily available for ICE-HK and ICE-IND, which turns an analysis based on intonational patterns into speculation. This is the first reason why topicalization here is understood in a wider sense, another is concerned with the information status of topicalized constituents.

In their publications, Birner and Ward repeatedly emphasise the constraint of preposed constituents having to be discourse-old:

> Felicitous preposing requires that the referent or denotation of the preposed constituent be anaphorically linked to the preceding discourse (see Reinhart 1981; Vallduví 1992). (Birner and Ward 1998, 32)

2 *Better* in (5) is written in capital letters by Ward and Birner to highlight the intonational focus. In (7), *he* represents the resumptive pronoun.

The study of information status has been largely shaped by Prince (1992), who created a matrix with different information statuses related to the discourse and the hearer. Information that is both discourse-old and hearer-old is called 'evoked,' while information that is old to the hearer but new in the discourse is referred to as 'unused.' Information that is entirely new is called 'brand-new,' whereas the combination of discourse-old and hearer-new is deemed impossible. According to Birner and Ward, only 'evoked' information can be topicalized. They claim that

> […] the constraint on preposing and postposing constructions is absolute (e.g. in preposing the preposed constituent must represent discourse-old information regardless of the status of the information represented by the rest of the sentence) […]. (Birner and Ward 2009, 1172)

The term 'discourse-old,' however, does not exclusively refer to an exact repetition of a previously mentioned entity. Instead, the relation holding between the preposed constituent and the entity in the preceding discourse may be one of "type/subtype, entity/attribute, part/whole, identity, etc." (Birner and Ward 1998, 32; Ward and Birner 2004, 159). Ward and Birner refer to the sum of the entities standing in any of these relations as 'posets,' i.e. partially ordered sets: "The notion of a poset subsumes both coreferential links, where the linking relation between the preposed constituent link and the corresponding poset is one of simple identity, and non-coreferential links, where the ordering relation is more complex" (Ward and Birner 2004, 159). In an example such as (5), for instance, it could be argued that the category 'sports' rather than the exact term 'baseball' is the topic (Birner and Ward 1998, 38). This 'givenness constraint' does not hold up in a close analysis of actual spoken language. Although evoked information clearly dominates in topicalized constituents, unused and sometimes even brand-new information may be topicalized as well under certain circumstances.[3] This finding is in line with Mesthrie (1992, 113), who identified tokens of topicalization containing both unused (8) and brand-new (9) information (emphasis added):

(8) *Your tablet* you took? (='Have you taken your tablets?')
(9) *Like a wild animal* you are.

The first example comes from a discourse without any prior mention of medicine, but it is implied by the way the question is asked that both discourse participants are aware of the fact that the addressee has to take medication. The second example, as Mesthrie (1992, 113) notes, "was the first statement of the day in a household, addressed to a cat trying to force open a window." Speaking

3 See also Chen (2003), who calls for a reassessment of the givenness constraint.

of hearer-old and discourse-old becomes a relative matter in this scenario, but believing the givenness constraint to hold in this case is certainly misguided, too.

Yet another constraint put forward by Birner and Ward (1998) concerns adverbials. In their framework, adverbials can only be considered to be instances of topicalization if they are "lexically governed by the matrix verb" (Birner and Ward 1998, 31). This criterion has also been applied in this study, although every case was treated individually and, when in doubt, a second rater was consulted in order to decide whether a sentence-initial adverbial is canonically positioned at the beginning of a clause or a case of topicalization. To sum up, the definition of topicalization in this paper is more liberal than Ward and Birner's and corresponds largely to Mesthrie's expanded concept of the term.[4]

A concept related to topicalization is that of topic-prominence, i.e. the general configuration of a language according to the topic-comment principle. For the present analysis, topic-prominence is relevant because in languages that are deemed to be topic-prominent, topicalization plays an important role and may even be entirely canonical. This way of structuring sentences might be transferred to English contact varieties, as it is part of the speakers' linguistic repertoire.

In order to establish the degree of topic-prominence in a language, Li and Thompson (1976) provide a list of characteristics they assume to be typical of topic-prominence. These characteristics and brief descriptions for them are given in Table 1. A full survey of these criteria goes beyond the scope of this paper. It should be noted, however, that any combination of the characteristics may be found in a language and that they may indeed be gradable to an extent (cf. Section 3).

As mentioned above, the relation between topicalization and topic-prominence lies in the fact that sentence-initial topics are less marked (or non-canonical) and more frequent to the degree of being the prevalent structure in a topic-prominent language's word order. This regularity of topics in sentence-initial position implies that other terminology might be preferable, since 'topicalization' inevitably gives the impression of a deliberate linguistic choice rather than denoting a regular state of things. As Plag (Plag 2003, 91; emphasis in the original) points out, "[d]erivatives in *-ion* denote events or results of processes." It should not go unmentioned that topicalization is often not 'required' even in a fairly strict SVX language such as English because subject and topic tend to overlap most of the time. Givón (1979, 210), for instance, estimates a discongruence between subject and topic in only

4 In addition to doing away with the givenness constraint, Mesthrie (1992) also predicts higher frequencies of topicalization in SAIE than in varieties spoken by white South Africans. Furthermore, he noted that different kinds of phrases may be topicalized and that the process interacts with embedding and other syntactic phenomena such as questions and negation (cf. Mesthrie 1992, 113–120).

Table 1: Characteristics of topic-prominent languages (based on Li and Thompson 1976).

	Characteristic	Description
(a)	Surface coding	Coding of the topic by position and/or morphological marking in topic-prominent languages
(b)	The passive construction	No or marginal passivization in topic-prominent languages
(c)	"Dummy" subjects	Absence of dummy subjects in topic-prominent languages
(d)	"Double subject"	Pervasive double-subject constructions in topic-prominent languages
(e)	Controlling co-reference	Topic controls co-referential constituent deletion in topic-prominent languages
(f)	V-final languages	Topic-prominent languages tend to be verb-final languages
(g)	Constraints on topic constituent	No constraints on what the topic may be in topic-prominent languages
(h)	Basicness of topic-comment sentences	Topic-comment structures are the basic structure in topic-prominent languages

10–20% of cases. An explanation for this is provided by Welke (Welke 1992, 57), who notes an "inherent thematicity of the subject,"[5] meaning that the (grammatical) subject has inherently topical characteristics. Further remarks on the concept of topic-prominence are provided in Section 4.

3 Methodology

The data for the present analysis come from the ICE components for Great Britain, Hong Kong, and India. Despite the fact that the age of some ICE components and the duration of corpus compilation imply a diachronic dimension (cf. Hundt 2015), the corpora still represent a formidable basis for comparing varieties (see Götz, this volume). Although topicalization may occur in written English, too, this study focuses on spoken language. As Schneider comments,

> [o]ne of its [=ICE's] strengths, quite clearly, is the size of its spoken components, given that oral performance is less constrained and less conservative than written styles, so this is where innovations are most likely to surface. (Schneider 2004, 247)

Thus, the conversation files of ICE-HK, ICE-IND, and, for reasons of comparison, ICE-GB were analysed.

If, as in this paper, a wider definition of topicalization is applied, the topicalized constituent may be any kind of phrasal constituent or a clause. Consequently, automated searches for topicalization (for instance by means of a regular expression), at this point, need to be restricted to a limited set of constituents. Only with a well-parsed and annotated corpus would an automated search for topicalization be feasible at all, and since the depth of annotation differs greatly between ICE components, all sections were read and tagged manually. The tags were then extracted and all relevant tokens were annotated for syntactic form, syntactic function, discourse function, information status and other criteria. Since the ICE corpora in their unaltered form have vastly different word counts, all annotation in the corpora, e.g. to indicate pauses or indigenous words, as well as any irrelevant contributions from speakers of other varieties were erased.[6] In a next step, the resulting figures were used to calculate the relative frequency of

5 English translation by the author of this paper; the original words in German are "inhärente Thematizität des Subjekts."

6 The word count without any annotation and mark-up amounts to roughly one million words per country/region in ICE; working with normalised frequencies was still preferred to minimize any remaining compilation bias.

topicalization tokens per 100,000 words. These numbers will be shown and discussed in the next section.

In addition to establishing the frequencies of topicalization in the varieties, typological profiles of the major contact languages of IndE and HKE with regard to topic-prominence (based on Li and Thompson 1976, cf. Section 2) were created. As mentioned above, a full description of all these criteria would go far beyond the scope of this paper, which is why the reader is referred to the original paper and subsequent discussions of Li and Thompson's framework.[7] Based on an in-depth study of relevant grammars, the characteristics given in Table 1 were rated according to the following scale:[8]

(A) ✓: the feature can be attested unambiguously;
(B) (✓): the feature is present to a limited extent (e.g. if there are some syntactic or pragmatic constraints);
(C) (×): the feature is present to a very limited extent (e.g. if it is highly marked or several constraints are in place);
(D) ×: the feature is absent;
(E) ?: none of the consulted sources explicitly or implicitly elaborated on the feature or the information was found to be ambivalent to a degree such that no rating could be decided on.

These ratings, of course, do not replace personal familiarity with the languages. However, they may serve to indicate tendencies. The selection of languages to which this procedure has been applied is explained and justified in the results section for each variety individually.

4 Case Studies

In this section, I present the results of the topic-prominence analysis as well as my findings from the corpus study. The first variety to be discussed is Indian English, i.e. the different forms of English spoken in the Indian subcontinent. Since India is a country of an enormous size with more than a billion inhabitants, speaking of Indian English as a monolithic variety is misleading and certainly misguided. Still, a common core of features seems to be shared by different regional subvarieties, one of which is topicalization (cf. Lange 2012).

7 See, for instance, Schlobinski and Schütze-Coburn (1992) and Junghare (1988).
8 Controlling co-reference, (e) in Table 1, was omitted because the majority of grammars did not provide any commentary on this phenomenon.

Table 2: Speaker percentages of six major languages of India in ICE-IND and in the population (adapted from Lange 2012, 83).

	Percentage of speakers in ICE-IND	Percentage of speakers in India's population
Hindi	5.81	41.03
Bengali (Bangla)	3.73	8.11
Telugu	7.47	7.19
Marathi	20.79	6.99
Tamil	12.86	5.91
Kannada	19.92	3.69

India's diversity is reflected in the large number of living languages in the country, reported to be at 447 by the *Ethnologue* (cf. Lewis et al. 2016). Table 2 provides information on the languages that were analysed with regard to topic-prominence in terms of their speaker numbers in India as well as the percentage of speakers who indicated these languages as their L1 in the ICE-IND metadata.

Taken together, the percentage of speakers in ICE in the second column of Table 2 accounts for 70.58% of speakers in the direct conversations. The selection of languages covers three Indo-Aryan (Hindi, Bangla and Marathi) as well as three Dravidian (Telugu, Tamil and Kannada) languages.[9] The results of the analysis of topic-prominence can be seen in Table 3.

Table 3: Topic-prominence features of six major Indo-Aryan and Dravidian languages.

	Hindi	Bangla	Marathi	Tamil	Telugu	Kannada
Surface coding	✓	✓	✓	(✓)	(✓)	(✓)
Lack of passives	×	(✓)	×	×	×	×
No dummy subjects	✓	✓	✓	✓	✓	✓
Double subjects	(✓)	(×)	(×)	(✓)	(✓)	(✓)
V-final language	✓	✓	✓	✓	✓	✓
No/few constraints	✓	(✓)	✓	✓	✓	✓
Basicness of TC	(✓)	(×)	(✓)	(×)	(×)	(×)

9 The sources that were consulted for the Indo-Aryan languages are Kachru (2006), Junghare (1988), Dasgupta (2003), Schmidt (2003), Shapiro (2003), Thompson (2012), Conners and Chacón (2015), Dhongde and Wali (2009) and Pandharipande (1997). Information about the Dravidian languages was taken from Lehmann (1989), Annamalai and Steever (1998), Kausen (2013), Shibatani (1999), Zvelebil (1990), Krishnamurti (1998), Subbarao (2004), Schiffman (1983), Jensen (1969) and Sridhar (1990).

The table shows a mixed distribution, with some languages meeting more of the criteria than others. The general tendency seems to be the same for the Indo-Aryan and the Dravidian languages, since they all show several traits of topic-prominence. Although one might hesitate to call these languages exclusively topic-prominent, Li and Thompson's classification of the Indo-Aryan languages as exclusively subject-prominent does not seem to hold, either. This is confirmed by previous studies, claiming, for instance, that "Indo-Aryan languages in general are more topic-prominent than subject-prominent" (Junghare 1988, 325). Junghare's study suggests influence from the Dravidian languages, meaning that the Indo-Aryan languages might have developed differently in terms of their word order configuration due to sharing the same space with Dravidian languages. Furthermore, Junghare (1988) claims that topic-prominence can indeed be graded; Bangla, according to her, appears to be less topic-prominent than either Hindi or Marathi. For the study at hand, the conclusion is sufficient that some of the major contact languages of Indian English are rather topic-prominent.

The second Asian variety analysed in this study is Hong Kong English. The most important language of Hong Kong as well as the city population's *lingua franca* is the Cantonese variety of Chinese. Setter et al. (Setter et al. 2010, 4) describe the language situation in Hong Kong as "trilingual and biliterate." While Cantonese, English and Mandarin are the three official spoken languages in Hong Kong, the "average Hongkonger" is able to write in Standard Chinese and English (Setter et al. 2010, 4). Census data reveals a consistent preference of Cantonese as "usual language," although English has become slightly more popular. After a decrease from 3.2 to 2.8% from 2001 to 2006, the 2011 Census indicated that 3.5% of Hong Kongers use English as their usual language. The percentages for Cantonese range from 89.2% in 2001 to 90.8% in 2006 and, finally, 89.5% in 2011. Putonghua, i.e. the colloquial variety of Mandarin Chinese, was named by 1.4% as their usual language in 2011. Other Chinese dialects account for 4% in the 2011 Census. These numbers very clearly show the persistent dominance of Cantonese in Hong Kong, and the lacking potential for English to develop into a language that truly serves as an identity carrier with significant intranational functions.

In terms of Li and Thompson's classification, Chinese and its varieties are often considered the prime examples of topic-prominence. Table 4 confirms this reputation, albeit not without some minor limitations.

The literature concerned with information structure in the Sinitic languages is almost unanimously in favour of calling them topic-prominent. Yip and Matthews (Yip and Matthews 2011, 84), for instance, believe that "the subject-predicate construction [within Chinese grammar] is a special case of topic-comment. Hence, if no other element is topicalized, the subject becomes the topic by default." This fits in with Givón's (1979) claim that subject and topic tend to overlap, and

Table 4: Topic-prominence features in Cantonese and Mandarin.[a]

	Cantonese	Mandarin
Surface coding	✓	✓
Lack of passives	×	×
No dummy subjects	✓	✓
Double subjects	✓	✓
V-final language	×	(×)
No/few constraints	✓	(✓)
Basicness of TC	✓	(✓)

[a]The sources consulted for the findings in this table are Li and Thompson (1981), Ross and Sheng Ma (2006), Shyu (2014), Lin (2001), Sun (2006), Cheng and Sybesma (2015), Hendricks (2003), Paul (2015), Yip and Matthews (2000), Yip and Matthews (2011), Killingley (1993), Setter et al. (2010) and Kausen (2013).

Cantonese as an SVO language is no exception to this rule (cf. Yip and Matthews 2011, 84; see also Yip and Matthews 2000, 115).

A contact hypothesis predicts that a high degree of topic-prominence results in increased frequencies of topicalization in the contact varieties. Judging by Tables 3 and 4, topicalization is therefore expected to be the most frequent in HKE and the least frequent in BrE. The results of the corpus analysis, as presented in Figure 1, reveal a very different picture.

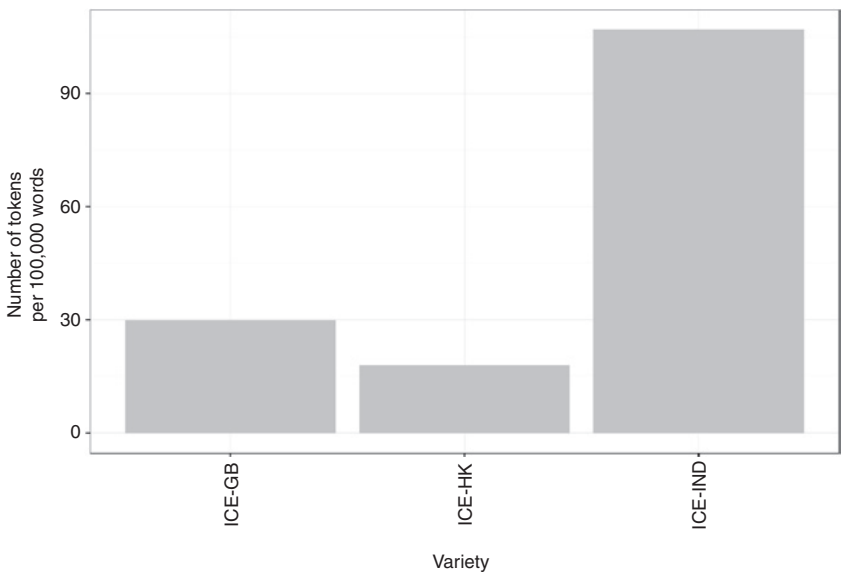

Figure 1: Distribution of topicalization in ICE-GB, ICE-HK and ICE-IND.

The figure shows that topicalization is most frequent in Indian English, with 106.91 instances of topicalization per 100,000 words. ICE-GB has the second-highest number with 29.99 per 100,000 and Hong Kong English has 17.91 cases of topicalization per 100,000 words. As explained in Section 2, these numbers are based on a rather liberal definition of topicalization, i.e. many adverbials were not excluded from the analysis and tokens did not have to contain discourse-old information to be counted. Examples (10–14) show different syntactic realizations of topicalization found in ICE-Hong Kong and ICE-India and reflect the great diversity of constituents that may be topicalized; the relevant constituent is always in italics.[10]

(10) Topicalized direct object/NP
B: And what you're going to have to eat
A: Now what do you want it's upto you
That you can decide
You want mutton you want fish you want chicken
<ICE-IND:S1A-003#100-103>

(11) Topicalized indirect object/NP
A: Friends are like more closer than parents in hostel
B: Parents means you can't tell each and everything
A: Yeah
B: Isn't it
A: *Friends* we can tell
<ICE-IND:S1A-054#202-206:1:A>

(12) Topicalized subject complement/NP
Z: It's kind of like half of A level
A: Uh we have the same
AS level yeah it is called
<ICE-HK:S1A-042#952-958>

(13) Topicalized obligatory adverbial/PP
A: Yah and they will run th they will run through the window yeah so *through the window* they can go inside the train
<ICE-HK:S1A-062#341:1:A>

(14) Topicalized direct object/clause
C: And I was the first man to go to driver and to scold him because see *whether you are a driver or not* first of all I asked him
<ICE-IND:S1A-017#16:1:C>

10 Any mark-up and annotation has been erased for better readability.

A chi-squared test indicates that the differences in terms of frequency are highly significant between the varieties ($\chi^2 = 90.328$, $df = 2$, p-value < 0.0005). Taking Figure 1 into account, this is not surprising; however, the reasons for this huge difference are in need of investigation. In the next section, I will discuss the contact-based approach and explore in which ways other factors influence the frequency and spread of topicalization.

5 Discussion

The results of the typological comparison are very different from what might be expected on the basis of an analysis of topic- and subject-prominence. With Cantonese being a highly topic-prominent language, the frequency of topicalization in HKE was expected to be much higher. The frequencies are, however, relatively low compared to IndE and BrE, which calls for a closer look at the various factors influencing the discourse-pragmatics and syntax of non-native varieties of English.

First, it needs to be stated that linguistic features on the level of syntax are transferred much less readily than features on the phonological or lexical level. Schneider's (Schneider 2007) Dynamic Model of Postcolonial Englishes, for instance, claims that varieties almost exclusively borrow phonological and lexical elements in their first developmental stages, i.e. in the phases of foundation and exonormative stabilization. At the time of the compilation of ICE-HK, i.e. pre-Handover in 1997, HKE can be assumed to have been in a transitional stage. It is unlikely that many features on any level were nativised, and information structuring strategies such as topicalization might only have occurred sporadically. Thus, an expected snowballing effect could not lead to a widespread diffusion of the feature. The snowballing effect describes the fact that "if many people use one particular linguistic variant, then the chances are higher that this variant will be selected for the stabilised variety" (Schneider 2007, 110). In a similar statement, Biewer (Biewer 2015, 104) notes that "[c]ontact phenomena may also be more lasting, as people will recognise them as having been used by others before."

A related factor influencing the frequency is what Thomason and Kaufman (1988) and Thomason (2001) termed 'intensity of contact.' This concept, later picked up by Matras (2009), relativises the temporal dimension of borrowing. Casual contact, according to Thomason and Kaufman (1988), encourages the borrowing of content words, but intense contact is required to result in changes in word order; or, in an extreme case, 'typological disruption' in Matras' framework (cf. Matras 2009, 156). 'Intensity,' though hard to define (cf. Thomason 2001, 66), encompasses different aspects. While power dynamics between the groups in contact play an

important role, the length of contact is an often underestimated factor. However, despite the undeniable importance of intensity of contact, interpreting data differences solely by looking at the length of contact between languages would drastically simplify matters. Matras (2009) acknowledges this problem and identifies a complex interplay of intensity of contact and several other aspects:

> A number of different factors are involved, including the degree of bilingualism and the roles that the languages have in various domains of social interaction, the degree of institutional support afforded to the languages (e.g. literacy, school instruction, media, language planning), and community attitudes. (Matras 2009, 156–157)

To a certain extent, these aspects can be accounted for in IndE and HKE and be put into relation to the intensity of contact. IndE is known to be one of the oldest postcolonial varieties, with its emergence dating back to the beginning of the 17th century (cf. Schneider 2007, 162). Hong Kong became a colony only after the first of the Opium Wars in 1841 (Schneider 2007, 133), with a notably late spread of English across the region. Further proof of the relatively slow development of the variety can be seen in the discussions at the turn of the millennium, when there was still much debate about whether HKE can even be considered a variety in its own right (cf., for instance, Bolton 2000). Although there is a functional divide in both India and Hong Kong, the range of functions of English in India goes far beyond those found in Hong Kong. Most importantly, perhaps, English is often used as a *lingua franca* in India, which is simply not necessary in a largely monolingual area such as Hong Kong. The conclusion to be drawn from this, then, is that contact was rather intense in parts of India, whereas there had been comparatively little language contact between English and Cantonese in Hong Kong up to the point the ICE data were collected. This can be seen in the fact that while there are undeniable structural differences between native varieties of English and HKE, they are often not very frequent and not as numerous.

The question may be asked which relation we can assume between canonical and non-canonical with regard to topicalization in HKE and IndE. In India, English has been present for 200 years longer than in Hong Kong, and transfer as well as a potential preference for topicalization in certain contexts from early on led to a spread of the feature across the country. Initially, it was probably highly marked despite the topic-prominent status of many Dravidian and Indo-Aryan languages. However, as time went on, a gradual pragmatic unmarking of the feature is likely. Although discourse functions can be assigned to any instance of topicalization, its high frequency suggests further unmarking in the future. It can now be called a prevalent part of Indian English, potentially making its way into a supraregional Indian English norm. In HKE, this

is certainly not the case. Although the feature is not absent from this variety, viewing it as part of a test scenario of common assumptions about borrowing and the diffusion of features as Schneider (Schneider 2007, 139) does is probably the best bet: "Hong Kong may become an interesting test case for the predictive implications of the Dynamic Model and the inherent power of the developmental dynamism which it describes." Thus, if 'canonical' and 'non-canonical' are defined as frequency-based terms, then topicalization must be declared a fairly canonical feature in IndE and a rather non-canonical feature in HKE. Consequently, the canonical/non-canonical divide appears to be linked to nativization, with the typology of contact languages only playing a role at later stages of a variety's development (if at all). In addition, the selection of features might be influenced by certain cultural predilections and processes of second-language acquisition.

6 Conclusion

Typologically speaking, Cantonese and Mandarin are considered highly topic-prominent languages. Several Indo-Aryan and Dravidian languages are, to a slightly lesser degree, topic-prominent as well. Thus, based on a contact hypothesis, one would expect ICE-HK to contain many more instances of topicalization, with 'more' referring to the relative amount in comparison to the other corpora. As the analysis of the conversation files in ICE has shown, this is evidently not the case: HKE features fewer cases than IndE, and fewer cases even than the British component of ICE.

Taking into consideration insights from research on language contact and variety development, this paper has shown that the frequency of topicalization appears to be, to a certain extent, linked to intensity of contact (in particular with regard to the temporal dimension) as well as the developmental stage of a variety. Although English is not considered to be an identity carrier in either India or Hong Kong, it has been a part of India's linguistic landscape for much longer than Hong Kong's. During its long history in the country, a relatively clear functional distinction between the Indian L1s and English has developed. As a result of intense language contact over hundreds of years, typological interference could occur and shape local forms of English much more intensely than in Hong Kong, where Cantonese has continually dominated in almost every sector. In particular, the role of English as a *lingua franca* in India meant actual communication in English in everyday contexts, which does not occur as frequently or intensely in Hong Kong.

The conclusion to be drawn from the analysis is that topicalization is not borrowed either quickly or easily, as HKE would otherwise feature it much more frequently. This finding is largely in line with Thomason and Kaufman's (1988), Thomason's (2001) and Matras' (2009) idea of intensity of contact. Although the precise mechanisms at work behind selecting topicalization might elude us based on an analysis of ICE, it seems likely that being part of the linguistic repertoire alone is not enough for topicalization to be used more frequently by speakers at an earlier stage of variety development.[11] Potentially, a highly exonormative orientation at the time of ICE-Hong Kong's compilation can thus explain why it shows so few instances of topicalization in relative comparison to ICE-India and ICE-Great Britain. Certain processes of second-language acquisition that may cause further spreading of the feature have not yet come into full effect, and the dominant role of Cantonese in Hong Kong also slowed down the diffusion of topicalization.

These findings are, to an extent, in line with Callies' (2009) suggestions on the transfer of information structure. Whether it is true that when learners' "L2 proficiency increases, native speakers of topic-prominent languages gradually increase the use of subject-prominent features in their L2 production" (Callies 2009, 91) cannot be substantiated or refuted without a good set of corpora featuring the same learners at different stages of their acquisition of English. However, the findings presented in this paper suggest that the following, somewhat open summary by Callies may indeed be linked to variety development: "Discourse structure and the pragmatic principles of information organization in the L1 may influence L2 acquisition in terms of transfer/overproduction […], or avoidance" (Callies 2009, 104). Avoidance, in this case, would refer to avoiding structures that appear to go against what is, for instance, taught in the classroom and therefore feel 'unnatural' or indeed highly non-canonical. Although the consequences of this are pure speculation, it might be added that naturally spoken English was much less 'accessible' at the time the ICE corpora for Hong Kong and India were compiled than it is today. Quick access to spontaneous spoken English entails quick access to many dialects and unsupervised language, which might have an influence on variety development.

For future research, analyses of topicalization in bigger corpora would be particularly valuable. Analysing topicalization on a larger scale and with more recent data would be needed in order to substantiate or reject the ideas put forward in this paper. Additionally, finding a way to extract tokens of topicalization with a regular expression, software or other automated means without

11 Diachronic corpora and/or longitudinal data would be needed to corroborate this claim.

missing a significant number of tokens would also be of great use to understand topicalization and similar phenomena better.

Works Cited

2011 Population Census. Main Report: Volume I. Hong Kong Special Administrative Region: Census and Statistics Department.

Annamalai, E. and Sanford B. Steever (1998). "Modern Tamil." Sanford B. Steever, ed. *The Dravidian Languages*. London: Routledge, 100–128.

Biewer, Carolin (2015). *South Pacific Englishes. A Sociolinguistic and Morphosyntactic Profile of Fiji English, Samoan English and Cook Islands English*. Amsterdam: John Benjamins.

Birner, Betty J. and Gregory Ward (1998). *Information Status and Noncanonical Word Order in English*. Amsterdam: John Benjamins.

Birner, Betty J. and Gregory Ward (2009). "Information Structure and Syntactic Structure." *Language and Linguistics Compass* 3.4: 1167–1187.

Bolton, Kingsley (2000). "The Sociolinguistics of Hong Kong and the Space for Hong Kong English." *World Englishes* 19.3: 265–285.

Callies, Marcus (2009). *Information Highlighting in Advanced Learner English. The Syntax-Pragmatics Interface in Second Language Acquisition*. Amsterdam: John Benjamins.

Chen, Rong (2003). *English Inversion. A Ground-before-Figure Construction*. Berlin: Mouton de Gruyter.

Cheng, Lisa L.-S. and Rynt Sybesma (2015). "Mandarin." Tibor Kiss and Artemis Alexiadou, eds. *Syntax: Theory and Analysis. An International Handbook*. Volume 3. Berlin: Mouton de Gruyter, 1518–1559.

Conners, Thomas J. and Dustin Chacón (2015). "Syntax." Anne Boyle David, Thomas J. Conners and Dustin Chacón, eds. *Descriptive Grammar of Bangla*. Berlin: Mouton de Gruyter, 249–302.

Dasgupta, Probal (2003). "Bangla." George Cardona and Dhanesh Jain, eds. *The Indo-Aryan Languages*. London: Routledge, 351–390.

Dhongde, Ramesh Vaman and Kashi Wali (2009). *Marathi*. Amsterdam: John Benjamins.

Givón, Talmy (1979). *On Understanding Grammar*. New York: Academic Press.

Hendricks, Henriëtte (2003). "Using Nouns for Reference Maintenance: A Seeming Contradiction in L2 Discourse." Anna Giacalone Ramat, ed. *Typology and Second Language Acquisition*. Berlin: Mouton de Gruyter, 291–326.

Hundt, Marianne (2015). "World Englishes." Douglas Biber and Randi Reppen, eds. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 381–400.

Jensen, Hans (1969). *Grammatik der kanaresischen Schriftsprache*. Leipzig: VEB Verlag Enzyklopädie.

Junghare, Indira Y. (1988). "Topic-Prominence and Zero NP-Anaphora." M. A. Jazayery and W. Winter, eds. *Languages & Cultures: Studies in Honour of Edgar C. Polomé*. Berlin: Mouton de Gruyter, 309–327.

Kachru, Yamuna (2006). *Hindi*. Amsterdam: John Benjamins.

Kausen, Ernst (2013). *Die Sprachfamilien der Welt. Teil 1: Europa und Asien*. Hamburg: Helmut Buske.

Killingley, Siew-Yue (1993). *Cantonese*. München: Lincom Europa.
Krishnamurti, Bh. (1998). "Telugu." Sanford B. Steever, ed. *The Dravidian Languages*. London: Routledge, 202–240.
Lange, Claudia (2012). *The Syntax of Spoken Indian English*. Amsterdam: John Benjamins.
Lehmann, Thomas (1989). *A Grammar of Modern Tamil*. Pondicherry: Pondicherry Institute of Linguistics and Culture.
Lewis, M. Paul, Gary F. Simons and Charles D. Fennig, eds. (2016). *Ethnologue: Languages of the World*. Nineteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.
Li, Charles N. and Sandra A. Thompson (1976). "Subject and Topic: A New Typology of Language." Charles N. Li, ed. *Subject and Topic*. New York: Academic Press, 457–489.
Li, Charles N. and Sandra A. Thompson (1981). *Mandarin Chinese. A Functional Reference Grammar*. Berkeley: University of California Press.
Lin, Hua (2001). *A Grammar of Mandarin Chinese*. München: Lincom Europa.
Matras, Yaron (2009). *Language Contact*. Cambridge: Cambridge University Press.
Mesthrie, Rajend (1992). *English in Language Shift. The History, Structure and Sociolinguistics of South African Indian English*. Cambridge: Cambridge University Press.
Mesthrie, Rajend (1997). "A Sociolinguistic Study of Topicalisation Phenomena in South African Black English." Edgar W. Schneider, ed. *Englishes Around the World Volume 2: Caribbean, Africa, Asia, Australasia. Studies in Honour of Manfred Görlach*. Amsterdam: John Benjamins, 119–140.
Pandharipande, Rajeshwari V. (1997). *Marathi*. London: Routledge.
Paul, Waltraud (2015). *New Perspectives on Chinese Syntax*. Berlin: Mouton de Gruyter.
Plag, Ingo (2003). *Word-Formation in English*. Cambridge: Cambridge University Press.
Prince, Ellen F. (1992). "The ZPG Letter: Subjects, Definiteness, and Information Status." William C. Mann and Sandra A. Thompson, eds. *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*. Amsterdam: John Benjamins, 295–326.
Reinhart, Tanya (1981). "Pragmatics and Linguistics: An Analysis of Sentence Topics." *Philosophica* 27.1: 53–94.
Ross, Claudia and Jing-heng Sheng Ma (2006). *Modern Mandarin Chinese Grammar. A Practical Guide*. London: Routledge.
Schiffman, Harold F. (1983). *A Reference Grammar of Spoken Kannada*. Seattle: University of Washington Press.
Schlobinski, Peter and Stephan Schütze-Coburn (1992). "On the Topic of Topic and Topic Continuity." *Linguistics* 30: 89–121.
Schmidt, Ruth Laila (2003). "Urdu." George Cardona and Dhanesh Jain, eds. *The Indo-Aryan Languages*. London: Routledge, 286–350.
Schneider, Edgar W. (2004). "How to Trace Structural Nativization: Particle Verbs in World Englishes." *World Englishes* 23.2: 227–249.
Schneider, Edgar W. (2007). *Postcolonial English. Varieties Around the World*. Cambridge: Cambridge University Press.
Setter, Jane, Cathy S. P. Wong and Brian H. S. Chan (2010). *Hong Kong English*. Edinburgh: Edinburgh University Press.
Shapiro, Michael C. (2003). "Hindi." George Cardona and Dhanesh Jain, eds. *The Indo-Aryan Languages*. London: Routledge, 250–285.
Shibatani, Masayoshi (1999). "Dative Subject Constructions Twenty-Two Years Later." *Studies in the Linguistic Sciences* 29.2: 45–76.

Shyu, Shu-Ing (2014). "Topic and Focus." C.-T. James Huang, Y.-H. Audrey Li and Andrew Simpson, eds. *The Handbook of Chinese Linguistics*. Malden, MA: Blackwell, 100–125.
Sridhar, S. N. (1990). *Kannada: Descriptive Grammar*. London: Routledge.
Subbarao, Karumuri V. (2004). "Non-nominative Subjects in Telugu." Peri Bashkararao and Karumuri V. Subbarao, eds. *Non-nominative Subjects*. Vol. 2. Amsterdam: John Benjamins, 161–196.
Sun, Chaofen (2006). *Chinese: A Linguistic Introduction*. Cambridge: Cambridge University Press.
The ICE Project. *International Corpus of English*. <http://ice-corpora.net/ice/index.htm>. (accessed September 2016).
Thomason, Sarah G. (2001). *Language Contact*. Washington, DC: Georgetown University Press.
Thomason, Sarah G. and Terrence Kaufman (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley and Los Angeles: University of California Press.
Thompson, Hanne-Ruth (2012). *Bengali*. Amsterdam: John Benjamins.
Vallduví, Enric (1992). *The Informational Component*. New York: Garland.
Ward, Gregory (1988). *The Semantics and Pragmatics of Preposing*. New York: Garland Publishing.
Ward, Gregory and Betty J. Birner (2001). "Discourse and Information Structure." Deborah Schiffrin, Deborah Tannen and Heidi E. Hamilton, eds. *The Handbook of Discourse Analysis*. Malden, MA: Blackwell, 119–137.
Ward, Gregory and Betty J. Birner (2004). "Information Structure and Non-canonical Syntax." Laurence R. Horn and Gregory Ward, eds. *The Handbook of Pragmatics*. Malden, MA: Blackwell, 153–174.
Ward, Gregory and Betty J. Birner (2011). "Discourse Effects and Word Order Variation." Claudia Maienborn, Klaus von Heusinger and Paul Portner, eds. *Semantics. An International Handbook of Natural Language Meaning*. Berlin: Mouton de Gruyter, 1934–1963.
Ward, Gregory, Betty J. Birner and Rodney Huddleston (2002). "Information Packaging." Rodney Huddleston and Geoffrey K. Pullum, eds. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press, 1363–1447.
Welke, Klaus (1992). *Funktionale Satzperspektive. Ansätze und Probleme der funktionalen Grammatik*. Münster: Nodius Publikationen.
Winkle, Claudia (2015). *Non-canonical Structures, They Use them Differently. Information Packaging in Spoken Varieties of English*. Freiburg: Universitätsbibliothek. <https://freidok.uni-freiburg.de/fedora/objects/freidok:10600/datastreams/FILE1/content> (January 29, 2017).
Yip, Virginia and Stephen Matthews (2000). *Basic Cantonese: A Grammar and Workbook*. London: Routledge.
Yip, Virginia and Stephen Matthews (2011). *Cantonese. A Comprehensive Grammar.* London: Routledge.
Zvelebil, Kamil V. (1990). *Dravidian Linguistics. An Introduction*. Pondicherry: Pondicherry Institute of Linguistics and Culture.